Vision Research 50 (2010) 2618–2626

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

The disembodied eye: Consequences of displacing perception from action

Roberta L. Klatzky [a,*], Bing Wu [a,b], George Stetten [b,c]

[a] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[b] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[c] Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA 15260, United States

ARTICLE INFO

Article history:
Received 19 May 2010
Received in revised form 19 July 2010

Keywords:
Action
Kinesthesis
Visualization
Localization
Anorthoscopic

ABSTRACT

In our research, people use actions to expose hidden targets as planar images displayed either *in situ* or *ex situ* (displaced remotely). We show that because *ex situ* viewing impedes relating actions to their perceptual consequences, it impairs localizing targets, including compensating for surface deformation, and directing movement toward them. Using a 3D analogue of anorthoscopic perception, we demonstrate that spatio-temporal integration of contiguous planar slices is possible when action and perception are co-located, but not when they are separated. *Ex situ* viewing precludes the formation of a spatial frame of reference that supports complex visualization from action.

1. Introduction

The field of motor control has been extensively concerned with mechanisms by which perception guides action, in such forms as pre-planning, closed-loop feedback, and feed-forward error correction (e.g., Wolpert & Flanagan, 2009). Perceptual support of action is apparent even in the rudimentary motor capabilities of infants (Bertenthal, Rose, & Bai, 1997). Conversely, action feeds into perception in various ways. Neisser (1976) famously used the term "perception/action cycle" to describe the iterative coupling by which actions change the perceived world and hence regulate further actions. Some motor commands are even produced specifically to effect perceptual consequences. For example, controlled head and eye movements are used to bring or maintain desired objects into the field of view, and research on touch has described dedicated "exploratory procedures" – patterns of movement that provide information about specific object properties (Lederman & Klatzky, 1987).

However perception and action interact, it is usually the case that people perceive directly in the space where they act. Take the paradigmatic case of dexterous object manipulation. Typically, the location of an object under manipulation is simultaneously where the eyes focus, the hands touch, and contact sounds are emitted. Displacement of perception from action can occur, however, as when a hand-held tool is used to manipulate an object. Visual focus is on the distal end of the tool, the source of feedback about contact, and away from the manipulatory movements of the hand itself. Tool users accommodate to this problem of perception/action decoupling quite well. Research suggests that the brain solves the problem of eye/hand separation by extending receptive fields of bimodal visual-haptic cells to encompass the tool as well as the hand (Iriki, Tanaka, & Iwamura, 1996). The rapidity of this adjustment suggests that intelligent species have developed evolutionary mechanisms that support highly proficient tool use.

Evolution is slow, however, and our contemporary world offers more taxing situations that decouple actions from their perceptual consequences. These situations transcend tool-based displacement and can result in visual displays that are far removed from the events they depict. One such situation is shown in Fig. 1. The left panel depicts a person positioning a transducer (sensing device) so as to acquire information about a hidden target. The device will be described further below, but its critical attribute is this: Much as a flashlight illuminates an object in the dark, the transducer not only detects the hidden target, but also exposes an image of it in 3D space, aligned with the target itself. The target is simultaneously placed in a frame of reference relative to the image, the action that exposes it, and the world. Through sensory-motor processing, all these frames become aligned in a representation of the common 3D space they occupy. The right panel shows a similar situation, but with a critical difference: Instead of the image of the target being displayed in its source location, it is displaced for viewing to a remote screen. The spatial frame of reference for the visible image is arbitrarily translated and rotated relative to the frames representing the action and the word. In effect, the operator's eye has become "disembodied".

Although the latter situation is commonplace in medical applications, relatively little is known about the consequences to

* Corresponding author. Fax: +1 412 268 2798.
E-mail address: klatzky@cmu.edu (R.L. Klatzky).

0042-6989/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2010.08.025

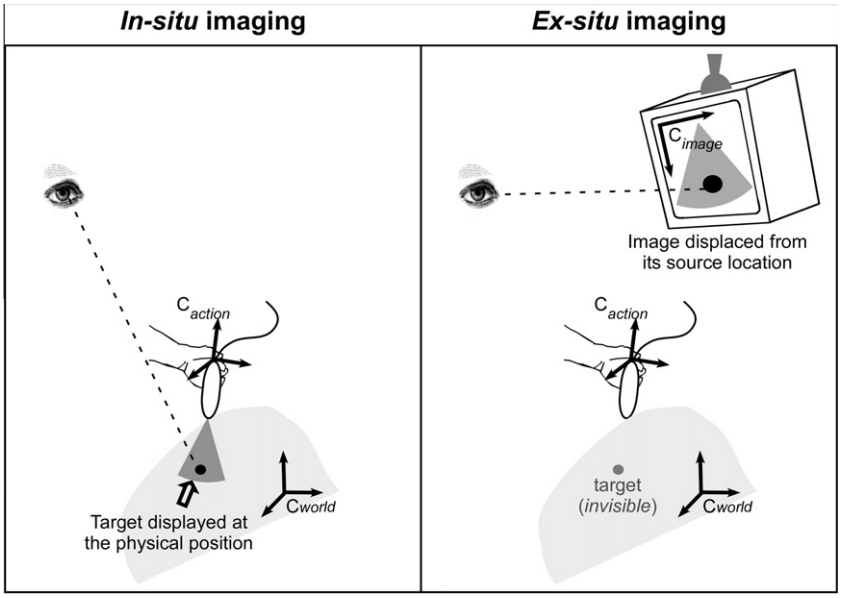

Fig. 1. Left panel: Imaging a hidden target in situ; the target is seen at its physical position in space. Right panel: Imaging a hidden target *ex situ*; the image containing the target is displaced to a remote screen.

perception and action. In the present review, we describe a research program that demonstrates, in a variety of tasks, the effects of displacing action from its visual base and hence disembodying the eye. The principal device we use in this research is a display originally developed for tomographic imaging of ultrasound data (Stetten & Chib, 2001). Ultrasound is a means of capturing a 2D map that shows the spatial locations of sound-reflecting objects in a scanned slice of the environment. Our research display presents the 2D plane of scanned data so that it appears as an image at its source in the 3D world. We use the term *in situ* for this form of presentation, which places a representation of data in full registration with the source location. As the user sweeps the display through the world, a succession of slices can be viewed *in situ*, potentially facilitating the user in reconstructing the 3D environment.

More generally, we will use the term *in situ* to describe not only the placement of the displayed data in the world, but also to denote the device that makes such a display possible (as in, an *in situ* device), and the experimental condition in which the data are displayed (as in, *in situ* viewing). These uses should be clear in context. Corresponding terms are used for the situation where the data slice is displaced to a remote screen, which we call displaying data *ex situ*.

We begin this review by describing conditions of *in situ* and *ex situ* viewing in detail. We then go on to describe a program of research in which these modalities were contrasted, in terms of utility for visualization and perceptually guided action. Our studies range from localization of occluded objects, to location-guided reaching, to accommodating for non-rigid occluders, and eventually to the visualization of increasingly complex 3D structures by exposing them one slice at a time. Across these tasks, we show that an *ex situ* display, which displaces the eye from the source data, has profound consequences for performance.

2. *In situ* and *ex situ* imaging

The *in situ* display developed for our research was initially designed for use with real-time tomographic images generated by an ultrasound scanner. *Tomographic* denotes images whose pixels represent data collected from discrete locations, generally organized as planar sections (*tomos* is Greek for *slice*), as opposed to *projection* images, whose pixels each represent data collected along an entire line. Stetten (2003) has developed a technique that visualizes ultrasound data in the following manner. A small display is embedded in the handle of an ultrasound transducer to produce an image, which is reflected by a half-silvered mirror, to produce a virtual counterpart of the tomographic image at its location in the scanned anatomy (see Fig. 2 for illustration). The technique is a form of augmented reality, as the half-silvered mirror illuminates and superimposes the ultrasound image of the inner tissue without occluding the direct view of the skin. By virtue of this illumination, the original device is called the "Sonic Flashlight." Our version of this device designed for research uses a mock transducer and a display that projects computer-generated data (Shelton, Wu, Klatzky, & Stetten, 2007). Like the original sonic flashlight, the projection through the mirror is seen in its source location in 3D space, together with any actual surfaces that may be present. Importantly, surfaces more proximal than the imaged data do not occlude the image itself.

Because light rays reflected from the mirror appear as if they come from the virtual image of the slice, the slice is perceived by the same mechanisms as would operate on a physical plane lying in space. That is, the viewer of the *in situ* plane has available full binocular depth cues, including convergence of the eyes and disparity between the left and right retinal images, as well as monocular cues including accommodation. The virtual slice is localized in 3D space by means of normal perceptual processes and merged with the viewer's body representation in egocentric space. It is on this basis that we denote the viewing of a virtual image through the device by the term *in situ*. Anyone who has ever had an ultrasound exam, however, will recognize that the *in situ* scenario is a departure from conventional clinical practice. Typically, the operator of an ultrasound scanner holds the transducer against the patient and gathers data, which are then sent to a monitor off to the side. This exemplifies the viewing condition we call *ex situ*.

A tomographic imaging display, whether real or virtual, creates a complex set of perceptual and cognitive processing demands on the user. Some demands arise because the system presents the world one 2D slice at a time. Given that the goal of the viewer is to represent the world in three dimensions, something that is

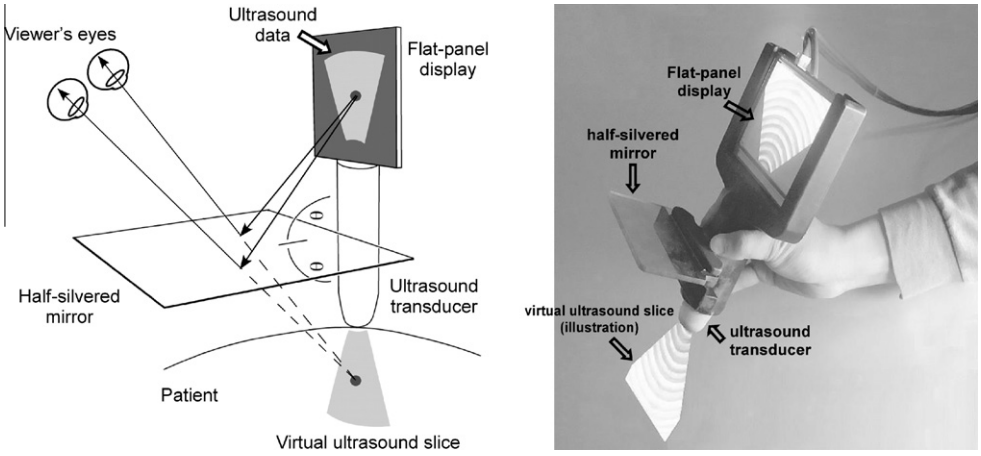

Fig. 2. Schematic and photograph of the *in situ* ultrasound imaging device. Through the half-silvered mirror, the image is projected as if the ultrasound "shines out" from the transducer and illuminates the scanned structures (from Wu, Klatzky, Shelton, and Stetten (2005), copyright IEEE, with permission).

necessary for guiding action, slice-by-slice display creates a demand for spatio-temporal integration. That is, successive planar views must be localized in relation to each other, in order for data seen in one to be related to data seen in another. This process is greatly facilitated if the locations of the slices can be perceptually situated in the 3D world of the user. Perceived spatial locations act as a "glue," providing a frame of reference in which data from different slices can be registered and hence forming the basis for their integration. The role of world localization is even more critical when actions must be directed toward the source images.

In the system that we utilize, a hand-held *transducer* senses *source data* in the world and converts the data to an *image*. In our tasks, the user's goal is to interpret the image as an object residing in peripersonal space (cf. conventional off-line use of recorded ultrasound images for medical diagnostics). To do so, he or she needs to know the position of the transducer as the source data are acquired, because transducer and data are mechanically linked. The fact that the transducer is hand-held means that cues to localization arise from multiple sources: The user can sense the transducer position by seeing it in 3D space, by kinesthetic cues from holding it, and from efferent copy of the motor commands that generated its current position. *In situ* imaging has the critical feature of projecting the image into the source data, creating a perceptually seamless whole in which vision, kinesthesis and action are all congruent.

The same cues to transducer location present with *in situ* imaging are also available with *ex situ* imaging. The user directs movement of the transducer and receives efference copy, is cued by kinesthesis as to its body-relative location, and – if viewing the transducer – can see it in space. However, the image is not in the same location as the source data, with the consequence that the user cannot look at the transducer and simultaneously see the image it acquires. On this basis, ex *situ* imaging creates a demand for integration across the space that displaces the viewed image from its perceptually–motorically cued origin. Whatever integrative process is required, it is not directly supported by perception; rather, it demands cognitive mediation. At a minimum, *ex situ* viewing requires mental translation between the observed image and the source of the data it portrays. Depending on placement of the display that portrays the image, and on the scaling of the image relative to the source data set (which in clinical practice, is essentially arbitrary), *ex situ* viewing may further invoke mental rotation and size adjustment, processes known to impose intense cognitive load (Bundesen & Larsen, 1975; Shepard & Metzler, 1971).

We have conducted a body of research concerned with the efficacy of imaging the 3D world through 2D slices, emphasizing the effects of the anomalous displacement of action imposed by *ex situ* viewing. In the remainder of this paper we will review these effects and their implications for perceptually guided action. We begin with localizing simple small targets and directing action to them, in the form of pointing and reaching with a needle.

3. *In situ* and *ex situ* imaging to access small targets

People's ability to localize and access small targets with *in situ* and *ex situ* imaging was assessed in a series of experiments (Chang, Amesur, Klatzky, Zajko, & Stetten, 2006; Wu, Klatzky, Shelton, & Stetten, 2005). The subject's task was to locate a target through ultrasound, and then to indicate its location by pointing to it or access it with a needle. In the Wu et al. study, a bead of 1 cm diameter was hidden in a tub of milky fluid and sonically imaged through a transducer held at the top of the tub. The target's perceived location was derived by having the subject point at the target, using a tracked pointing device, from three different locations around the rim of the tub. For each pair of pointing vectors, an intersection was determined, and the centroid of the three intersection locations was used to determine where the subject spatially localized the target (Fig. 3a). In another task, the subject directed a needle to the imaged target, along a trajectory perpendicular to the image plane (Fig. 3b). A similar needle insertion task was implemented by Chang et al. but the targets were tubes (simulated veins) in a blue-gel medium. The task was to thread a needle into the tube at an unspecified location along its length, and the principal measure was response time.

In the Wu et al. study, and also in that by Chang et al., the tasks were performed either *in situ*, by means of the sonic flashlight display, or *ex situ*, in which case the image was projected on a remote screen. The results of these experiments showed clear advantages for *in situ* visualization. In the Chang et al. study measuring response time, naive subjects and experienced ultrasound users were found to be faster to access targets localized *in situ*. With the most difficult target, the *ex situ* device resulted in a 50% increase in access time (11 s, cf. 7 s for *in situ*). In the Wu et al. study, where localization of targets was assessed from the pointing task, *in situ* viewing was found to be as accurate as when the same targets were directly seen in an unfilled tub (labeled "direct vision" in Fig. 3). In comparison, use of the *ex situ* display led to a systematic under-estimation error (bias) in the visualized depth of the targets, on the order of 1 cm.

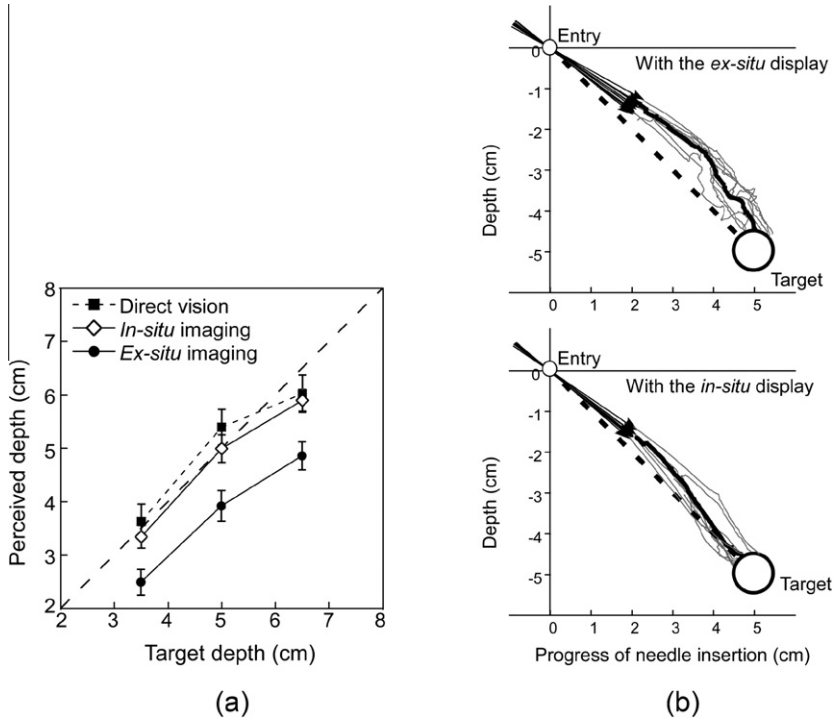

Fig. 3. Results of Wu et al. (2005) for localizing and accessing small targets with the *in situ* and *ex situ* displays. (a) Mean judged depth of targets as a function of the physical depth; in the direct-vision condition the participant looked at the bead within a tank devoid of fluid. (b) Trajectories of successful needle insertions performed by a naïve subject. Bold black lines depict the average across the trials. Arrows depict the needle's initial aiming direction as positioned by the subject at the beginning of each trial (from Wu, Klatzky, Shelton, and Stetten (2005), copyright IEEE, with permission).

Wu et al. further showed that the representation of target location, whether formed from perception (*in situ*) or cognitive mediation (*ex situ*), predicted the actions directed toward the targets. When subjects tried to access the ultrasound-visualized target with a needle, they aimed directly at targets visualized *in situ* and, since the targets were localized accurately, they proceeded along a straight trajectory (Fig. 3b, bottom). When visualizing *ex situ*, however, subjects' aiming trajectories were consistent with the localization bias inferred from the pointing task: They initially aimed at too shallow a depth, and then corrected when the needle could be observed approaching and penetrating the target slice. As a result, they followed a curved trajectory toward the target. This tendency led to greater variability (noise) in the trajectory followed under *ex situ* guidance, as is evident in the example in Fig. 3b (top).

In short, these studies show errors in localizing targets when their images are displaced, in the form of both bias and noise. From a practical standpoint, these errors are clinically significant and may contribute to the difficulties that have been documented in using ultrasound for vascular access (Keenan, 2002). From a scientific standpoint, errors provide insight into the process by which the representation of location is formed. In particular, the systematic under-estimation bias observed with *ex situ* localization may be related to visual distance errors in reduced-cue contexts (e.g., Gogel, 1969). Deformability of the surface contacted by the transducer may also contribute, as we discuss next.

4. *In situ* and *ex situ* localization: compensating for deformability

As was noted above, localizing a target from an *ex situ* device is a complex interaction involving kinesthesis and motor efference, which provide feedback about the position of the transducer as it is pressed against the target; vision, which perceives the image from the transducer at a displaced location; and cognition, which must bridge the gap between target location and viewed image. Now add to this complexity a non-rigid world. We did so by introducing the possibility of deformation in the surface that supports the transducer.

In a tomographic data set, the position of a target is provided in transducer-relative coordinates. Clinical ultrasound specifically conveys the depth of the target relative to the transducer tip by displaying a metric scale in the image. Now consider what happens if the transducer touches a non-rigid surface, beneath which lies a fixed target. As force is applied to the transducer, it deforms the surface and approaches the target. As force is released and the surface recovers, the transducer recedes from the target. The movements of the transducer relative to the target are portrayed in the image plane by shifts in the target's vertical location, which is higher (shallower) in the image plane as force is applied and lower (deeper) as force is released. But in our scenario, since the target is fixed in the world, these changes in its vertical location in the image are misleading; they reflect transducer-relative location rather than world coordinates. To perceive the fixed location of the target in the world, the user must somehow calibrate the movements of the transducer and adjust for the consequences in the image.

Calibrating transducer movements is not difficult when an *in situ* device is used (Fig. 4, left). Because the user's view of the image also includes the transducer, its displacement is visible, and more importantly, the shifts in the target's vertical location on the image are seen as resulting from that displacement. As the image moves down with greater force applied to the transducer, the target moves up within the image by the same amount. The target is perceptually localized in the very same world coordinates, regardless of the force applied to the transducer and its resulting penetration into the occluding surface. The task becomes more

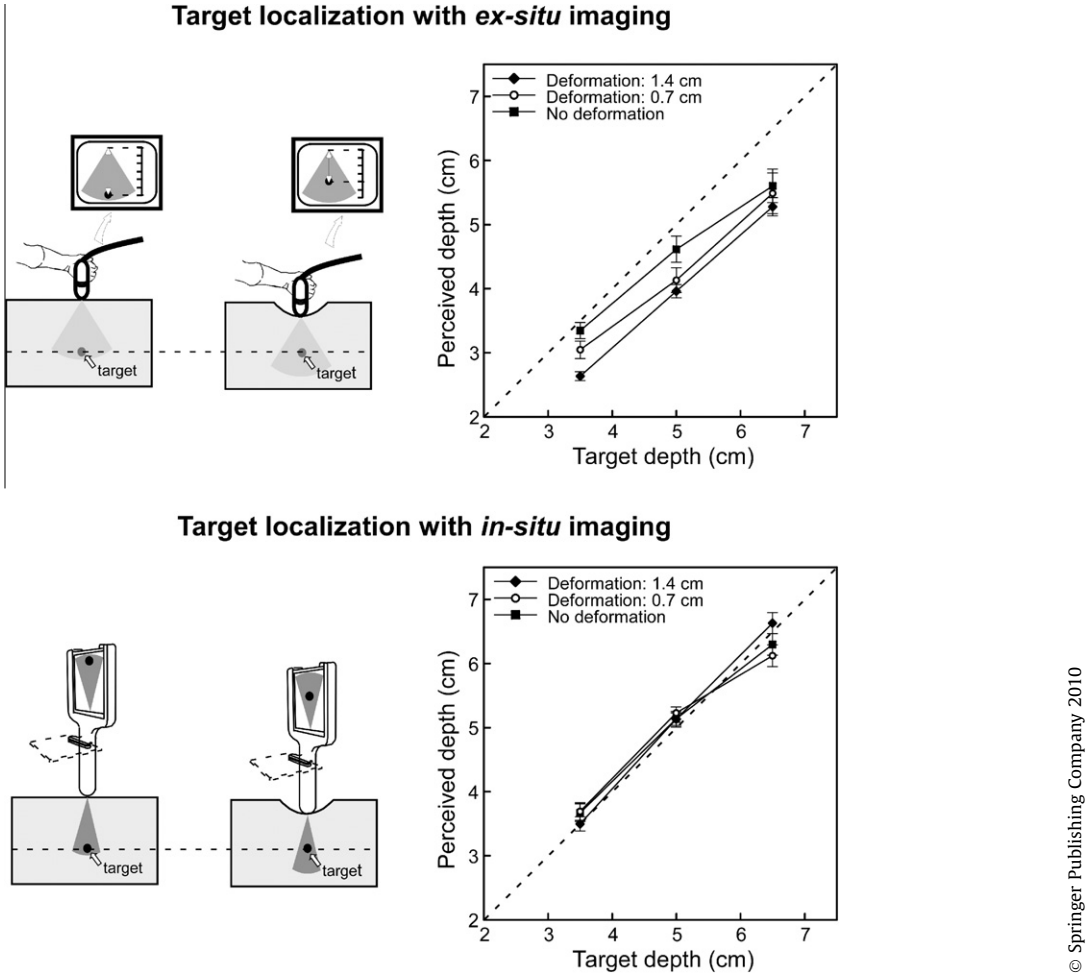

Fig. 4. Left: Localizing a target in a deformable medium with the *ex situ* and *in situ* displays. Right: Mean judged depth is plotted as a function of the target depth and the amount of deformation (from Wu, Klatzky, Shelton, and Stetten (2008), with permission).

difficult, however, if the image is viewed *ex situ*. Now, in order to compensate for the displacement of the surface and the resulting change in the target's vertical location on the image plane, the user must take into account feedback about transducer position.

In a series of experiments (Wu, Klatzky, Shelton, & Stetten, 2008), we showed that *ex situ* visualization is vulnerable to shifts in the transducer-relative target location as the transducer penetrates a deformable surface (Fig. 4, right). We performed a variant of the localization task described above, where subjects used ultrasound to sense a target hidden in a tub and then pointed to it from the rim of the tub. In the center of the tub cover was placed a step-down inset with a controlled depth. This was covered by a patterned, deformable "skin," beneath which were elastic bands. The result was an occluding surface that produced resisting force as it was penetrated. Subjects were instructed to exert sufficient force to make the transducer "bottom-out" in the step-down area. They then pointed to the target from the rim of the tub, which lay above the indentation.

We hypothesized that objective indentation of the transducer and the resisting force would affect the judgments of users of the *ex situ* display, but not those using the *in situ* display, which gives a "ground truth" (or better said, "depth truth") target location in perceived 3D space, independently of transducer position. As predicted, *ex situ* viewers were clearly affected by the deformability of the surface. This occurred despite the fact that they had strong cues to transducer indentation. They could watch the transducer as it penetrated the step-down area and came to a stop. The depth of the step-down was cued by perceptual cues to depth, including stereo vision and the deformation of the pattern on the surface of the "skin"; in addition, kinesthesis and efferent cues provided redundant cues to transducer location. However, when the subjects looked away from the transducer at an image of the target, its localization was affected by both the depth of the step-down and the resisting force. These essentially independent effects were mediated by perceived transducer indentation. Under-estimation of indentation led to under-estimation of target depth. In addition, higher force led subjects to assume the transducer was more deeply indented; as a result, as the resisting force increased, targets were estimated as deeper.

5. *In situ* and *ex situ* representation of the metric 3D world

Up to this point, we have presented research dealing with the effects of displacing target images on localizing the physical target in 3D space. The targets were essentially point objects at a particular depth and hence could be imaged in a single slice. We now turn to a more complex process, namely, integrating across slices in space and time in order to produce a representation of objects that extend into 3D.

A phenomenon that has been of interest to perception researchers for well over a century (Helmholtz, 1867; Zöllner, 1862) is

people's ability to form a representation of an object when the contours are exposed over time. The term *anorthoscopic perception* has been used to denote the integration of an object over successive views. The basic paradigm uses a 2D aperture to present an observer with a series of piecemeal views of an object on a picture plane. The views may be implemented by moving the object behind a slit or other opening in an otherwise opaque field, or a viewing region may move over the object. The depicted object itself may be stationary or in motion. Over many variations in the paradigm, there is broad agreement that the information in temporally and spatially continuous partial exposures is sufficient to lead to an object representation that transcends the limited aperture available at any point in time, one that often encompasses the whole object (e.g., Fendrich, Rieger, & Heinze, 2005).

Tomographic imaging, real or virtual, offers a 3D analogue to anorthoscopic perception. Instead of a moving slit that passes over an object on a picture plane, a moving plane passes through an object situated in 3D space. Just as researchers before us have asked whether successive slit-views of an object can be integrated to produce a representation of its planar form, we asked whether successive slice-views of an object can be integrated to produce a representation of its 3D structure. And, just as others have investigated the distortions that can result from slit viewing over time, we can ask whether there are distortions that arise from slice viewing over time. Our technology, in which the slice is passed through space under the control of an observer, and the resulting data are displayed *in situ* or remotely, *ex situ*, leads us to ask a further question: Does the ability to integrate slices into 3D forms benefit from *in situ* viewing? Our analysis suggests that this is likely to be the case.

To investigate people's ability to integrate 2D slices into 3D objects, we chose what seems initially to be a very simple task (Wu, Klatzky, & Stetten, 2010): The object to be visualized was a virtual rod with a circular cross section of 1 cm diameter (Fig. 5). The rod was rendered in space within an occluding box (31 cm long × 17 cm wide × 22 cm high). We assigned axes to the box such that *x*, *y*, and *z* were width, height and length, respectively. Two angular parameters then described the rod: its *pitch*, or up/down tilt around the *x*-axis as in Fig. 5, and its *yaw*, or left/right tilt around the *y*-axis. Subjects were asked to run a mock transducer over the centerline of the box along its length (rigid runners enforced this centering). They then were asked to report one or both of these parameters – pitch alone, yaw alone, or both pitch and yaw – by adjusting a response rod in 3D space to match the tilt of the virtual rod.

As they ran the transducer along the box, at any moment the subjects saw a planar image containing a roughly circular cross section of the rod. If the rod tilted in the pitch direction only, then as the transducer scanned along the box, this cross section would rise or fall on the vertical axis of successive images (corresponding to the *y*-axis of the box) while remaining centered horizontally. Whether the cross section rose or fell would depend on the direction in which the transducer was moved relative to the pitch direction; for example, a rod pitched downward toward the subject would cause the cross section to move down as the transducer moved toward the subject. Similarly, if the rod was yawed only, the cross section would pass rightward or leftward (depending on direction of the user's arm motion and the direction of yaw), while remaining centered vertically. Importantly, the rate of up/down movement when pitch varied, and similarly, the rate of right/left movement when yaw varied, would depend jointly on the magnitude of the tilt angle and the rate of speed of the transducer. If a user moved the transducer at constant speed, greater tilt angles would cause the cross section to show greater rates of displacement over successive images. If a user moved faster, this would also cause greater rates of displacement for a given tilt angle.

Now consider what happens when both pitch and yaw vary. The cross section of the rod on the image plane undergoes simultaneous movement on both the *x*- and *y*-axis as the transducer is moved. The magnitudes of the tilt angles, together with the rate of movement of the transducer, determine the speed at which the visible cross section moves in the *x*- and *y*-axis directions over successive images. These speeds may be quite different if the tilt

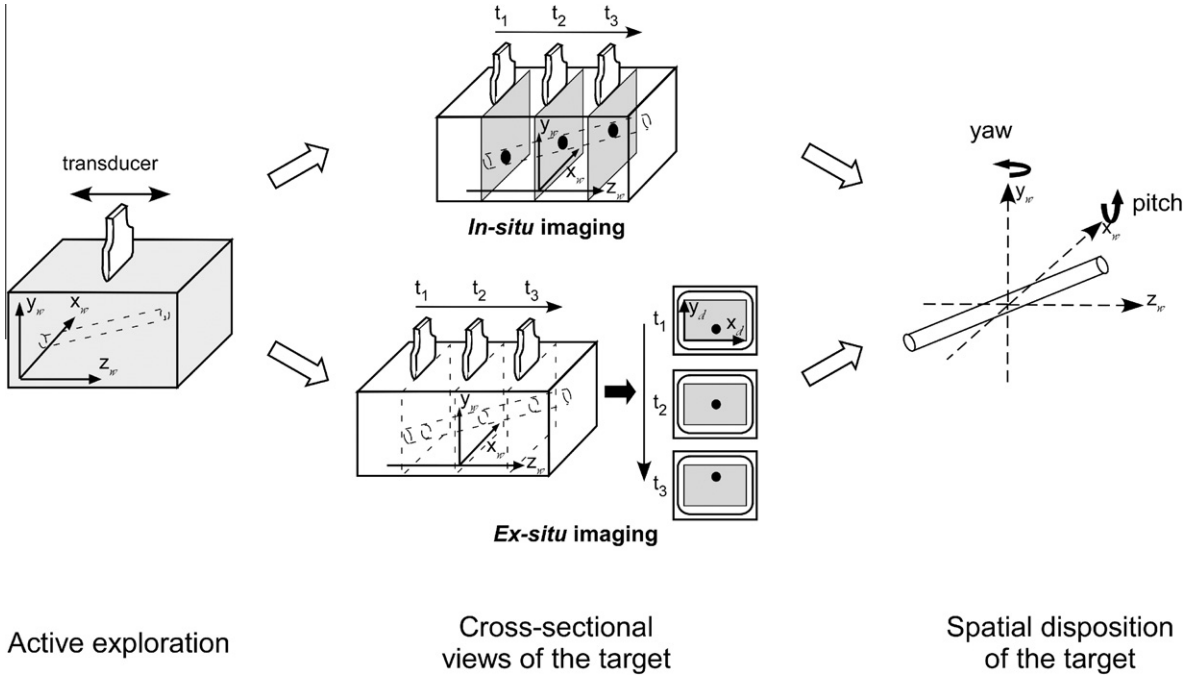

Fig. 5. Illustration of the 3D anorthoscopic perception task. Participants were asked to explore a hidden, virtual rod with an *in situ* or *ex situ* imaging device, exposing it as a successive sequence of cross sectional images, and then to estimate its orientation. The subscripts "*w*" and "*d*" denote world and display coordinates, respectively (from Wu, Klatzky, and Stetten (2010), with permission).

angles are discrepant in magnitude. For example, the cross section of a steeply pitched rod with small yaw will rise or fall quickly over successive images, while the right-left movement goes more slowly.

We reasoned that this task would be facilitated by *in situ* viewing, because the images produced by the hand as it exposes the rod are co-located with the hand movements themselves. As the arm moves the transducer along the box, the location of the current cross section is conveyed by visual cues and kinesthetic cues from arm movement. Simultaneously, the cross section of the rod is observed *in situ*, lying at the location of the virtual rod itself within the box. Keep in mind, however, that although perception of any one cross section is congruent across vision and action, only one slice can be seen at a time. Thus to visualize the rod as a whole, there remains a highly non-trivial process of integrating across its successive cross sections. We predicted, nonetheless, that the coherence of the cross sections in perception–action space would allow subjects to construct a representation of the rod in 3D space under *in situ* viewing.

Ex situ viewing, however, presents a more complex processing situation, as the cross sectional images are displaced from their source within the 3D world defined by the arm and the box. In an attempt to reduce demands, we placed the screen displaying the images so that it was parallel to the plane of the imaged slice, avoiding the requirement of mental rotation. Moreover, the scale of the displaced image was 1:1 with the rendered rod and hence matched to the *in situ* condition, eliminating demands for mental rescaling. Still, the displaced viewing means that the location of the slice at any time must be sensed by kinesthesis and motor efference, and the content of the viewed image must mentally be mapped into the perceived spatial location, in order for 3D localization of the cross section to occur. Further, the rate of transducer movement must be perceived, in order for users to relate the visible position changes in the cross section of the rod over a series of images to a spatial extent of arm movement. Only then can the image-relative movement be understood in terms of the rod's pitch and/or yaw angle.

Our results were clear in showing that *in situ* imaging of planar slices was sufficient to build up a highly accurate metric representation of pitch and/or yaw. Under *ex situ* viewing as well, subjects achieved a reasonable level of accuracy as long as pitch or yaw varied in isolation. However, a substantial number of *ex situ* errors were observed in which magnitude was correct, but direction of tilt was reversed (e.g., reporting the rod was pitched 45° toward the observer when it was pitched 45° away). This error pattern suggests that subjects were not visualizing the rod as an integral object in space, but rather were inferring tilt from 2D image cues and tagging direction separately.

Even more telling were the errors in the *ex situ* condition when both pitch and yaw varied together, particularly when their magnitudes were discrepant, as shown in Fig. 6. In this case, a pictorial cue – the amount the cross section translated within the

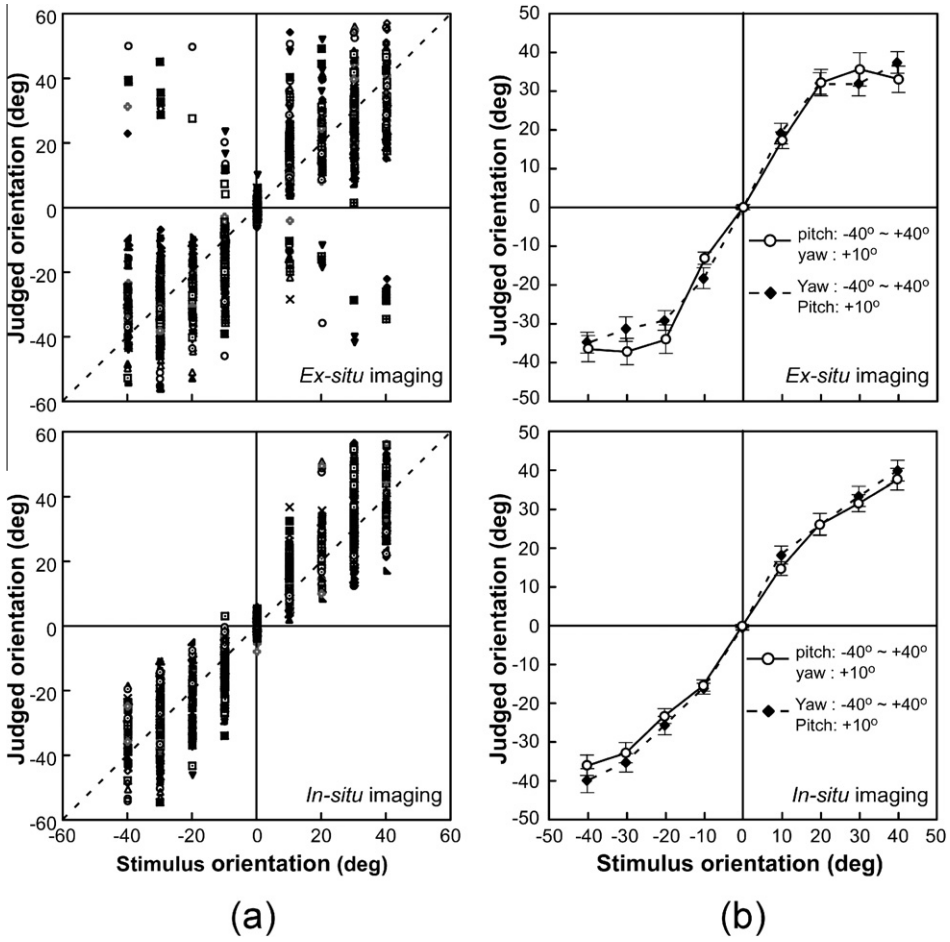

Fig. 6. Results of an experiment where subjects simultaneously judged pitch and yaw of a rod. Two sets of stimuli were constructed by varying yaw levels from −40° to 40°, while the rod's pitch was set at 10° or 30°, and vice versa. Data from all trials (pitch and yaw judgments, both stimulus sets) are pooled and plotted in (a) as a function of the stimulus attribute varying across the ±40° range. (b) The average judgments with the two displays. The open and filled symbols correspond to judgments of pitch and yaw across the ±40° range, when the other orientation (yaw or pitch) was fixed at 10°. Error bars represent 1 s.e.m. (from Wu, Klatzky, and Stetten (2010), with permission).

image – dominated responses. For example, when the rod was yawed to a large magnitude but pitch-tilt was small, the cross sections traveled quickly along the *x*-axis of the image plane before exhausting the visible length of the rod, leaving only a short region of space in which pitch data were visible. If the viewer assessed orientation from the observed translation in the mage, instead of constructing a representation of the rod in 3D space, this situation led to over-estimation of pitch. Directly corresponding effects of pitch on yaw were observed.

On the whole, these results confirmed our hypotheses: A 3D version of anorthoscopic perception, where planar slices are integrated into 3D objects, can be achieved with metric accuracy. The constraints are that we used a simple object, and success required that the slice images be co-located with the virtual object within the space of exploration.

6. *In situ* and *ex situ* construction of complex objects

Given the metric accuracy with which a simple object, the rod, could be visualized *in situ*, we set a higher goal: to visualize the 3D structure of more complex objects by sweeping through them to expose planar slices. Experiments in this series (Wu, Klatzky, & Stetten, 2008a; in preparation) have used unfamiliar forms constructed of interconnected line segments (see Fig. 7, left for illustration), unfamiliar closed forms lying in 2 or 3 dimensions, and familiar alphanumeric symbols. In all cases, the stimuli were virtual objects that were rendered within the same box as used in the previous experiment with the rod, and the subject's task was to visualize the object by sweeping the transducer along the box. The success of visualization was tested either with a subsequent same/different judgment or, for the familiar alphanumeric stimuli, by the accuracy of identification.

It should not be a surprise to the reader at this point that *in situ* viewing facilitates the process of visualization, relative to the displacement of the image that constitutes *ex situ* viewing. In the experiment illustrated in Fig. 7, a 2×2 ANOVA (display type as a between-subject factor; pattern complexity manipulated within-subject) showed that the time to construct patterns was greater for *ex situ* than *in situ* viewing ($F(1, 18) = 7.97$, $p = 0.011$), and that pattern complexity independently affected construction time ($F(1, 18) = 117.801$, $p < 0.001$). (For the interaction, $F < 1$.) The complexity effect indicates that the time to visualize a pattern increases as it has more segments that are exposed over multiple slices, rather than appearing entirely within a single slice (like the vertical bar in an I, compared to the crossbar). A segment lying oblique to the scanning axis (e.g., the middle section of a **Z** scanned from top to bottom) is particularly difficult to construct, because its cross sections are not only acquired over time, but they also lie at different locations in the image plane over successive slices.

7. Action to perception: the disembodied eye

The research reviewed in this paper shows clear disadvantages for performance in a variety of tasks when perception is displaced from the space of action; that is, when the eye is disembodied. The first set of studies using hidden targets showed that displaying a viewed object outside of its spatial context impairs the ability to localize it by pointing, move toward it along a straight-line trajectory, and compensate for deformation of an occluding surface. The second set of studies, using a 3D analogue of anorthoscopic perception, showed that the ability to construct a mental representation of a target by exploring it in slices was impeded when the slice images were displaced from the field of exploration.

As was noted in the introduction, disembodying the eye creates demands for new forms of integration that relate actions to their perceptual consequences across space. In the version of our task with *ex situ* viewing, people use actions to expose targets as a series of planar images. The actions produce motor efference and kinesthetic feedback, and they may also be viewed by the person performing the task. However, the visual feedback about action is not available at the same time as the resulting image of the source data, because the image and action reside at different spatial locations.

This disruption of the normal relation between action and its perceptual consequences may seem innocuous, but it has far-reaching consequences. It leads to distortions of perceived spatial location within single images. When multiple, temporally linked images that represent spatially contiguous source data are presented, the perception/action dislocation appears to preclude the formation of a common spatial frame of reference that allows them to be integrated into a visualized whole. The data from *in situ* visualization show clearly that there are circumstances where spatio-temporal integration is possible, as long as it operates over a coherent spatial frame. The data from *ex situ* visualization show that disembodying the eye makes this integration impossible.

Many questions arise from these studies that merit further research. An important issue concerns learning: Is it possible for long-term users to compensate for the separation of action-produced images and their display? Our own studies give a pessimistic

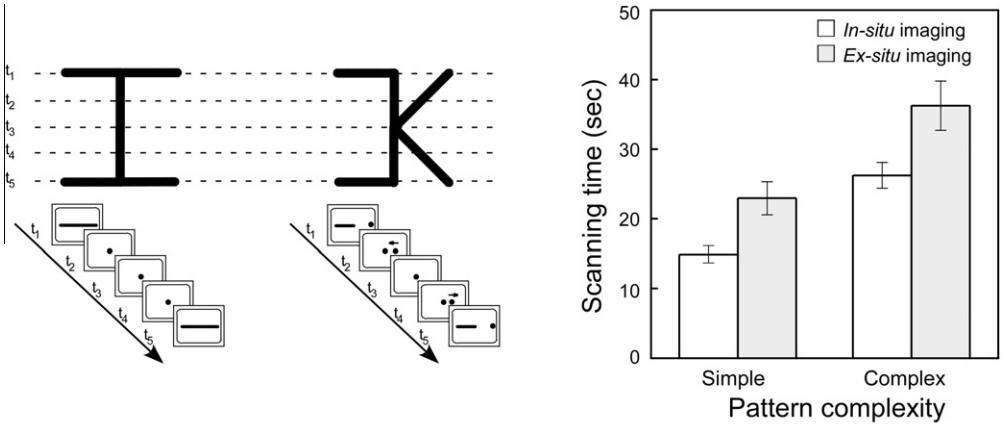

Fig. 7. Mental construction of complex objects from cross sections. (a) Examples show the sections of two patterns (left: simple; right: complex) observed at five locations and time points. (b) Time to scan the pattern in preparation for a same/different judgment, by imaging display and complexity.

answer to this question. We found that extensive practice on a single target did not generalize even to approaching the same target from a new position (Klatzky, Wu, Shelton, & Stetten, 2008). Our experiments suggest that what is learned is a response calibration, that is, a mapping from a specific representation of a target location to an intended insertion response (Wu, Klatzky, & Stetten, 2008b). However, in these learning studies, practice was restricted to a relatively small number of trials. We are greatly interested in pursuing the effects of long-term learning, particularly in clinical contexts where experts may be found.

Our previous studies with experienced practitioners indicate that despite extensive experience with *ex situ* displays, after minimal training they readily adapt to *in situ* imaging. Experienced nurses rated *in situ* guidance of simulated catheter placement easier than the conventional *ex situ* procedure (Chang et al., 2006), and both experienced radiologists (Amesur et al., 2009) and intravenous team nurses (Wang et al., 2009) demonstrated equivalent effectiveness of *in situ* and *ex situ* displays for guiding catheter placement in clinical trials. These studies testify to people's remarkable ability to learn novel perception/action couplings, but they also reaffirm our general conclusion that actions are best supported when they are spatially united with the perceptual system.

Acknowledgments

This research was supported by National Institutes of Health Grants # R01EB00860-03 and R21EB007721-01 and National Science Foundation Grant # 0308096.

References

Amesur, N., Wang, D., Chang, W., Weiser, D., Klatzky, R., Shuka, G., et al. (2009). Peripherally inserted central catheter placement in the interventional radiology suite using the sonic flashlight. *Journal of Vascular and Interventional Radiology, 20*, 1380–1383.

Bertenthal, B. I., Rose, J. L., & Bai, D. L. (1997). Perception–action coupling in the development of visual control of posture. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 1631–1643.

Bundesen, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 214–220.

Chang, W., Amesur, N., Klatzky, R., Zajko, A., & Stetten, G. (2006). Vascular access: Comparison of US guidance with the sonic flashlight and conventional US in phantoms. *Radiology, 241*, 771–779.

Fendrich, R., Rieger, J. W., & Heinze, H.-J. (2005). The effect of retinal stabilization on anorthoscopic percepts under free-viewing conditions. *Vision Research, 45*, 567–582.

Gogel, W. (1969). The sensing of retinal size. *Vision Research, 9*, 3–24.

Helmholtz, H. V. (1867). *Handbuch der physiologischen Optik*. Hamburg: Voss.

Iriki, A., Tanaka, M., & Iwamura, Y. (1996). Coding of modified body schema during tool use by macaque postcentral neurons. *NeuroReport, 7*, 2325–2330.

Keenan, S. P. (2002). Use of ultrasound to place central lines. *Journal of Critical Care, 17*, 126–137.

Klatzky, R. L., Wu, B., Shelton, D., & Stetten, G. (2008). Effectiveness of augmented-reality visualization vs. cognitive mediation for learning actions in near space. *ACM Transactions on Applied Perception, 5*(1), 1–23. <http://doi.acm.org/10.1145.1279640.1279641>.

Lederman, S. J., & Klatzky, R. L. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology, 19*, 342–368.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco: W.H. Freeman.

Shelton, D., Wu, B., Klatzky, R., & Stetten, K. (2007). Design and calibration of a virtual tomographic reflection system. In IEEE international symposium on biomedical imaging, ISBI, Washington, D.C., April 2007.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701–703.

Stetten, G. (2003). System and method for location-merging of real-time tomographic slice images with human vision. US Patent No. 6599247. Issue date: July 29, 2003.

Stetten, G., & Chib, V. (2001). Overlaying ultrasound images on direct vision. *Journal of Ultrasound in Medicine, 20*, 235–240.

Wang, D., Amesur, N., Shukla, G., Bayless, A., Weiser, D., Scharl, A., et al. (2009). Peripherally inserted central catheter placement with the sonic flashlight: Initial clinical trial by nurses. *Journal of Ultrasound in Medicine, 28*, 651–656.

Wolpert, D. M., & Flanagan, J. R. (2009). Forward models. In T. Bayne, A. Cleermans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 294–296). Oxford, England: Oxford University Press.

Wu, B., Klatzky, R. L., & Stetten, G. (2008a). Exploring here, seeing where? Visualization with in-situ vs. ex-situ viewing. In Annual meeting of the Vision Science Society, May 2008.

Wu, B., Klatzky, R. L., Shelton, D., & Stetten, G. (2005). Psychophysical evaluation of in-situ ultrasound visualization. *IEEE Transactions on Visualization and Computer Graphics, 11*(6), 684–693.

Wu, B., Klatzky, R. L., Shelton, D., & Stetten, G. (2008). Mental concatenation of perceptually and cognitively specified depth to represent locations in near space. *Experimental Brain Research, 184*, 295–305.

Wu, B., Klatzky, R. L., & Stetten, G. (2008b). Learning to reach to locations encoded from imaging displays. *Spatial Cognition & Computation, 88*, 333–356.

Wu, B., Klatzky, R. L., & Stetten, G. (2010). Visualizing 3D objects from 2D cross sectional images displayed in-situ versus ex-situ. *Journal of Experimental Psychology: Applied, 16*(1), 45–59.

Zöllner, F. (1862). Über eine neue art anorthoskopischer zerrbilder. *Annalen der Physik und Chemie, 117*, 477–484.